Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
Информатики и радиоэлектроники

УДК 004.021

Якубович
Андрей, Владимирович

Методы сбора и анализа данных в социальных сетях

АВТОРЕФЕРАТ
На соискание степени магистра технических наук
по специальности 1-40 80 02 "Системный анализ, управление и обработка информации"

Научный руководитель
Шилин Леонид Юрьевич
д.т.н., профессор

Минск 2020

ВВЕДЕНИЕ

Социальные сети в сети интернет являются отражением структуры и динамики современного общества. Объём трафика в социальных сетях настолько велик, что составляет более половины трафика всей сети интернет. Люди проводят в них все больше времени: там они развлекаются, общаются, ищут нужную информацию, советуются, знакомятся и устанавливают деловые отношения

По данным агентства Mordor Intelligence, по состоянию на 2017 год мировой рынок услуг анализа публикаций в социальных сетях оценивался в 3 млрд долларов США и, согласно прогнозам, достигнет 16 млрд долларов США к 2023 году [44]. Аудитория подобных продуктов - в первую очередь крупные и средние B2C-компании. Их задачи - своевременное выявление негативных отзывов, бенчмаркетинг, поиск лидеров мнений, евангелистов и антагонистов бренда, сравнение с конкурентами, а также оценка успешности рекламных кампаний. Помимо этого, сервисы востребованы у маркетинговых и PR-агентств, которые используют их для внешней независимой оценки результатов своей работы и оказывают с их помощью услуги мониторинга. Перспективной нишей для развития услуг мониторинга открытых источников является рынок информационной безопасности. Анализ упоминаний позволяет своевременно обнаруживать информационные атаки, их источники, а инструментарий работы с социальными сетями возможность оперативной реакции для минимизации ущерба. Это говорит об актуальности исследований методов сбора и анализа дынных социальных сетей.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель данной работы – изучить популярные методы сбора и анализа данных социальных сетей. Разработать алгоритм решения задачи выявления ботов в социальной сети Вконтакте. Провести текстовый анализ данных трафика социальной сети.

Объектом исследования являются социальные сети. Предметом исследования – алгоритмы сбора, анализа и визуализации данных социальных сетей.

Для достижения поставленной цели были поставлены следующие задачи:

− выбор социальных сетей, представляющих наибольший интерес для анализа и его обоснование;

− выбор наиболее подходящих технологий из существующих на сегодняшний день для построения интерфейса информационной системы и анализа полученных данных;

− реализация необходимых алгоритмов получения, анализа и визуализации данных;

− сравнение результатов разработанного алгоритма с наиболее популярными методами выявления ботов по параметрам точности, полноты и F-меры.

− мониторинг трафика социальной сети и построение модели городского пространства.

Для реализации алгоритмов анализа данных наиболее подходящим признан язык программирования Python. Сбор данных осуществляется многопоточной системой парсинга страниц социальной сети. Модули парсеров написаны на языке программирования C#. Панель администрирования реализована по технологии web api.

Часть данных, необходимых, в частности, для построения модели городского пространства, получена с использованием облачного сервиса Google BigQuery и данных, предоставляемых социальной сетью Twitter для обучения нейронных сетей.

СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из четырёх глав.

В первой главе рассматриваются популярный методы сбора и анализа данных социальных сетей. Описываются основные подходы к анализу социального графа пользователей. Определена проблема анализа современных социальных сетей и определен выбор алгоритмов для её решения в данной диссертации.

Вторая глава рассматривает вопросы проектирования и программной реализации информационной системы анализа сбора и анализа данных социальных сетей. Определены требования к системе мониторинга. Спроектирована панель администрирования и многопоточная система парсинга страниц социальной сети Вконтакте.

Третья глава посвящена разработке алгоритма выявления ботов в социальной сети на основе метода выявления пересекающихся сообществ. Предоставлены экспериментальные результаты предложенного алгоритма. Произведено сравнение по параметрам полноты, точности и F-меры предложенного алгоритма с алгоритмами максимизации модулярности и алгоритмом CESNA.

В четвёртой главе проводится исследование городского пространства на основе трафика социальной сети Twitter. Проведён сбор сообщений пользователей связанного с происходящим в городе Индианаполис спортивным мероприятием, проведена фильтрация, кодирование и соотнесение текста сообщений с географическими объектами. Построена модель города Индианаполис по ключевым элементам городского пространства.

ЗАКЛЮЧЕНИЕ

В ходе магистерской диссертации был проведен анализ существующих методов сбора и анализа данных в социальных сетях. Особое внимание уделено методам, основанных на теории графов и методах текстового анализа. Выявлены недостатки популярных методов анализа социальных графов. Обоснована необходимость разработки новых методов анализа графа ближайшего окружения пользователей социальных сетей.

Разработан быстрый и масштабируемый алгоритм выделения пересекающихся сообществ в социальных сетях с атрибутами вершин, не требующий полноты данных атрибутов и устойчивый к их частичному отсутствию. На основе разработанного алгоритма предложен метод выявления ботов, основанный на качественном анализе сообществ графа ближайшего окружения пользователя. Метод был опробован на двух выборках управляемых и автоматических ботов социальной сети Вконтакте. Эксперимент продемонстрировал высокие значения точности, полноты и F1-меры обнаружения ботов. Применение разработанного метода может существенно повысить сложность создания правдоподобных аккаунтов ботов.

Проведён текстовый анализ данных социальной сети Twitter для построения модели городского пространства на основе трафика посетителей Суперкубка 2012 года.

Были спроектирована программные модели для передачи и обработки данных, механизм API для взаимодействия модулей системы. Создано программное обеспечение для многопоточного сбора данных из социальной сети. Собраны тестовые выборки графа пользователей из социальной сети ВКонтакте, которые были использованы для экспериментальной оценки качества разработанных методов и алгоритмов анализа графа ближайшего окружения пользователя социальной сети.

Благодаря слабосвязанной архитектуре отдельные модули системы могут применяться для создания приложений мониторинга и анализа данных социальных медиа и могут использоваться в компаниях, занимающихся разработкой автоматизированных систем мониторинга с минимальным количеством доработок.

На основе данной диссертации опубликованы три тезиса на научные конференции.

Таким образом, цель и поставленные задачи были достигнуты в полной мере.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1] Якубович, А.В. Анализ влияния социальных сетей на мобильность пользователей / Информационные технологии и системы: материалы международной научной конференции – Минск, 2019. – С. 84 - 85.

[2] Якубович, А.В. Идентификация участников общественных мероприятий на основе данных социальных сетей / Информационные технологии и системы: материалы международной научной конференции – Минск, 2018. – С. 280 - 281.

[3] Якубович, А.В. Системный анализ городской среды посредством мониторинга социальной сети Twitter / Информационные технологии и системы: материалы международной научной конференции – Минск, 2020.